We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

3

Relational Analysis for Clustering Consensus

Mustapha Lebbah, Younès Bennani and Nistor Grozavu
LIPN - UMR 7030 CNRS, Université Paris 13,
99, avenue Jean-Baptiste Clément
93430 Villetaneuse.
e-mail:firstname.secondname@lipn.univ-paris13.fr
France

Hamid Benhadda
Thales Land & Joint 160, Bld de Valmy - BP 82
92700 - Colombes cedex.
e-mail:Hamid.BENHADDA@fr.thalesgroup.com
France

1. Introduction

One of the most used techniques among many others in the data mining field is the clustering. The aim of this technique is to synthetize and summarize huge amounts of data by splitting it into small and homogenous clusters such that the data (observations) inside the same cluster are more similar to each other than to the observations inside the other clusters. This definition assumes that there exists a well defined clustering quality measure that quantifies how much homogeneous are the obtained clusters. The aim of this chapter is to expose an original approach to merge different partitions, related to the same data set, which are obtained either by applying different clustering techniques either by the same clustering technique with different parameters. Fusing partitions has been broadly studied and has been given several names, depending on different scientific fields, like machine learning or bioinformatics (Dudoit & Fridlyand, 2003; Kim & Lee, 2007; Monti et al., 2003). Among these names we can quote: consensus clustering, clustering aggregation, clustering combination, fusion of clustering, ...etc. Several studies (Frossyniotis et al., 2002; Minaei-Bidgoli et al., 2004; Strehl & Ghosh, 2002; Topchy et al., 2004; 2005) have pioneered clustering data sets as a new branch of the conventional clustering methodology. In (Topchy et al., 2004) the authors propose a probabilistic formalism of clustering concensus using a finite mixture of multinomial distributions in a space of clustering. The approach proposed in (Frossyniotis et al., 2002) is designed for combining runs of clustering algorithms with the same number of clusters. In (Strehl & Ghosh, 2002) the authors proposed combiners based on a hyper-graph model to solve the cluster fusion problem. The authors discuss two manners of consensus clustering: (1) Feature Distributed Clustering (FDC): a set of clustering are obtained from partial view of variables using all observations (2) Object-Distributed Clustering (ODC): with this technique the ensemble clustering has limited to subset of observation with access to all variables. The

authors provide three techniques (CSPA[1], HGPA[2], MCLA[3]), but indicate that HGPA delivers poor scores for the both data sets used in this chapter. Our work is in FDC category. In (Azimi et al., 2007) authors propose a modification of k-means for clustering a multiple runs of k-means. It's named intelligent k-means, which is especially defined for clustering ensembles. All these models assume that the correct number of clusters is given as parameter of model. In (Gionis et al., 2007) the authors give a formulation of ensemble clustering titled clustering aggregation, which does not require a number of clusters. The authors give a nice review of algorithm dedicated to ensemble clustering.
In this chapter, we offer a representation of consensus clustering as a set of new variables characterizing the observations. This leads to a formulation of the fusion problem as categorical clustering problem. We propose to use *Relational Analysis (RA)* as consensus method for unsupervised learning.The concensus clustering is provided as solution of the minimization of the objective function for a given consensus clustering. The main idea, shared with other algorithm is : If many clustering algorithms assign two observations in the same cluster, it will not benefit to consensus clustering to split these observations.
There are several advantages of RA consensus function: first we have low computational complexity, and second ability to deal with huge data set. Another purpose of our algorithm is not to neglect the weak clustering result. Often in the ensemble/aggreation/fusion clustering we combine only the best results. Given observations and m clustering result proposed with categorical variables, the purpose is to produce a single clustering that agree as much as possible with all results of clustering algorithms. The algorithm we propose for the problem of consensus clustering takes advantage of statistical formulation, (Benhadda & Marcotorchino, 2007).

Relational Analysis as clustering fusion can be applied in various settings. Multiple runs of clustering algorithm, like self-organizing map, generate a new variable space, which is significantly better than pure or normalized variable space. Therefore, running a simple clustering algorithm on generated variable space can provide the consensus cluster significantly better than pure observations. In this chapter we present another features of our framework:

- *Clustering categorical variable*: Consensus clustering provides a natural method for clustering categorical data.
- *Determining the correct number of clusters*: The formulation we propose don't require as parameter the number of clusters. The only parameter needed by RA is the similarity threshold.
- *Clustering mixed data* : the clustering fusion method can be particularly effective in the cases where data are defined over heterogeneous variables that contain incomparable values. We consider in this chapter a particular case, that we deal with continuous and categorical variables. In such cases the data set can be divided vertically into sets of homogeneous variables. Thus we apply an appropriate clustering algorithm and then combine the individual clustering into single clustering using categorical data clustering method.

[1] Cluster-based Similarity Partitioning Algorithm
[2] HyperGraph Partitioning Algorithm
[3] Meta-Clustering Algorithm

The rest of the chapter is structured as follows: In section 2 we describe in detail the proposed model for consensus clustering. In section 3 we present a special case of global fusion based on self-organizing map. In section 4 we present experiments on public data set.

2. Relational analysis framework

Relational analysis theory is a mathematical data analysis approach with a broad application field. It was initiated and developed by (Marcotorchino & Michaud, 1978) at the IBM's European Center of Applied Mathematics (ECAM) by the end of the seventies. This technique uses the concept of "pairwise comparisons" which has been introduced in the statistical literature by the end of the thirties, through the work of (Kendall & Smith, 1940). Nevertheless the concept which has inspired the previous authors, dates of 1785 based upon some works of the "marquis de Condorcet" (Condorcet, 1785), related to "voting theory". In a general way, *Relational Analysis* makes it possible to model and solve problems whose general formulation can be stated as : *Seeking a particular relation $\mathcal{R}$ which fits "as well as possible" single (or several) given relations* $\left(\mathcal{R}^1, \mathcal{R}^2, \ldots \mathcal{R}^m\right)$.

Unlike the existing clustering techniques, RA methodology does not need necessarily, neither to do sampling to be able to get results in a raisonable computing time, nor to fix arbitrarily the number of clusters that could be hidden in the data.
The principle of "pairwise comparisons" consists in transforming, each variable V measured on N objects into a $N \times N$ squared matrix C representing the similarity, with regards to variable V, between the N^2 couples of objects. An illustration of the "pairwise comparison principle" can be found in (Benhadda et al., 2007).

2.1 Relational analysis clustering methodology

To cluster a data set $\mathcal{P}$ composed of n observations $(O_1, O_2, ..., O_n)$ described by m variables $(V^1, V^2, ..., V^m)$, we firstly start by transforming each column V^k into a relational matrix C^k with general term $c^k_{ii'}$ defined by:

$$c^k_{ii'} = \begin{cases} 1 & \text{if } O_i \text{ and } O_{i'} \text{ have the same modality of variable } V^k \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

This term representing the similarity between the observations O_i and $O_{i'}$, with respect to variable V^k. Once all the m matrices C^k had been built up, we construct a global relational matrix C called "Condorcet's matrix" of general term $c_{ii'}$ representing the global similarity of O_i and $O_{i'}$ with respect to the whole set of the m variables: $c_{ii'} = \sum_{k=1}^{m} c^k_{ii'}$. This global similarity has the so called "self maximal similarity property defined by: $c_{ii'} \leq \mathcal{M}_{ii'} \forall O_i, O_{i'}$, where $\mathcal{M}_{ii'} = Min(c_{ii}, c_{i'i'})$ is the "maximum possible similarity" between the two observations O_i and $O_{i'}$.

Using the global similarity $c_{ii'}$ and the "maximum possible similarity" $\mathcal{M}_{ii'}$ between O_i and $O_{i'}$, we define their dissimilarity $\bar{c}_{ii'}$ as the complement of their global similarity to their "maximum possible similarity":

$$\bar{c}_{ii'} \quad = \quad \mathcal{M}_{ii'} - c_{ii'} \quad (2)$$

Two observations will be, a priori, in the same cluster of the final expected partition as soon as their similarity will be greater than their dissimilarity i.e.: $c_{ii'} \geq \overline{c}_{ii'}$. The required final partition will be represented by a $N \times N$ binary squared matrix X with general term $x_{ii'}$ defined as follows:

$$x_{ii'} = \begin{cases} 1 & \text{if } O_i \text{ and } O_{i'} \text{ are in the same cluster} \\ & \text{of the final partition} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This partition will be obtained by maximizing the Condorcet's criterion $\mathcal{C}(X)$ defined hereafter:

$$\mathcal{C}(X) = \sum_{i=1}^{n} \sum_{i'=1}^{n} (c_{ii'} x_{ii'} + \overline{c}_{ii'} \overline{x}_{ii'})$$

where:

$$\overline{x}_{ii'} \quad = \quad 1 - x_{ii'} \tag{4}$$

Using the expressions (2) and (4), the criterion $\mathcal{C}(X)$ can be rewritten as:

$$\mathcal{C}(X) = \sum_{i=1}^{n} \sum_{i'=1}^{n} (2c_{ii'} - \mathcal{M}_{ii'}) x_{ii'} + \sum_{i=1}^{n} \sum_{i'=1}^{n} \overline{c}_{ii'} \tag{5}$$

As the second member of the sum of expression (5) is a constant, we deduce that maximizing the Condorcet's criterion is equivalent to maximizing the following criterion $\mathcal{C}'(X)$

$$\mathcal{C}'(X) = \sum_{i=1}^{n} \sum_{i'=1}^{n} \left(c_{ii'} - \frac{\mathcal{M}_{ii'}}{2} \right) x_{ii'}$$

The cost function of the criterion $\mathcal{C}'(X)$ will be positive when the similarity $c_{ii'}$ between two observations O_i and $O_{i'}$ is greater or equal to half of their "possible maximal similarity". This condition is sometimes very difficult to reach, especially when the number of variables (or descriptors) is very high compared to the number of observations i.e. $m >> n$, this is usually the case when the data set to be clustered is a set of documents. In that case, the number of clusters of the final partition will be so high that it could deprive the clustering task of its interest for practical purpose. As the goal of the clustering task is to summarize the amount of data into simpler structures, to avoid this problem, a solution consists in relaxing the cost function related to the clustering criterion. To reach that goal it is sufficient to replace the coefficient $1/2$ of $\mathcal{M}_{ii'}$ by a parameter α such that $0 < \alpha < 1/2$. The new formulation of the criterion $\mathcal{C}'(X)$ will be then:

$$\mathcal{C}'(X) = \sum_{i=1}^{n} \sum_{i'=1}^{n} (c_{ii'} - \alpha \times \mathcal{M}_{ii'}) \, x_{ii'}$$

Thus, the mathematical formulation of the relational analysis clustering problem is:

$$\max_{X} \mathcal{C}'(X)$$

under the constraints:

$$\begin{cases} x_{ii'} \in \{0,1\} & \forall \, (O_i, O_{i'}) \in \mathcal{P}^2 & \text{(binarity)} \\ x_{ii} = 1 & \forall \, O_i \in \mathcal{P} & \text{(reflexivity)} \\ x_{ii'} - x_{i'i} = 0 & \forall \, (O_i, O_{i'}) \in \mathcal{P}^2 & \text{(symmetry)} \\ x_{ii'} + x_{i'i''} - x_{ii''} \leq 1 & \forall \, (O_i, O_{i'}, O_{i''}) \in \mathcal{P}^3 & \text{(transitivity)} \end{cases}$$

2.2 The RA heuristic

The exact solution of the problem above can be obtained by linear programming techniques when the studied population is relatively small (few hundreds). But, in practice, the data set size can often exceed hundreds of thousands or millions of observations. This situation leads to use heuristics, to get the "best" and closest partition to the exact one, in reasonable time processing. We give below the description of the heuristic which was used by the relational analysis methodology in the eighties.

Phase 1

This step consists in intializing the clustering process by building a first partition. To build up this first partition we construct progressively its clusters according the operations described bellow:

1. Initialization: we take randomly a first observation which constitutes the first cluster of the unknown partition
2. We take an observation $O_i \in \mathcal{P}$, and compute its link $\mathcal{L}_{i\mathcal{V}}$ (expression 6) with all the existing clusters $\mathcal{V}$.

$$\mathcal{L}_{i\mathcal{V}} = \sum_{i' \in \mathcal{V}} \mathcal{L}_{ii'} \tag{6}$$

   where the link $\mathcal{L}_{ii'}$ between O_i and $O_{i'}$:

$$\mathcal{L}_{ii'} = c_{ii'} - \alpha \times \mathcal{M}_{ii'} \tag{7}$$

   This observation is assigned to the cluster which has the biggest strictly positive link with. If all the links are negative, then we create a new cluster to put in this new observation.
3. **Repeat** this process until all observations of population $\mathcal{P}$ had been assigned to a cluster.

Phase 2

At the end of the first step, we obtain a partition with a number of clusters[4].

1. **Merging two clusters:** We take, now, the clusters one after another and we compute the link $\mathcal{L}_{\mathcal{V}\mathcal{V}'}$ (expression 8) of each cluster $\mathcal{V}$ with all the others $\mathcal{V}'$.

$$\mathcal{L}_{\mathcal{V}\mathcal{V}'} = \mathcal{A}_{\mathcal{V}\mathcal{V}'} - \alpha \times \mathcal{M}_{\mathcal{V}\mathcal{V}'} \tag{8}$$

   where the agreement $\mathcal{A}_{\mathcal{V}\mathcal{V}'}$ between the two clusters:

$$\mathcal{A}_{\mathcal{V}\mathcal{V}'} = \sum_{i \in \mathcal{V}} \sum_{i' \in \mathcal{V}'} c_{ii'}.$$

   The disagreement $\overline{\mathcal{A}}_{\mathcal{V}\mathcal{V}'}$ between the two clusters is:

$$\overline{\mathcal{A}}_{\mathcal{V}\mathcal{V}'} = \sum_{i \in \mathcal{V}} \sum_{i' \in \mathcal{V}'} \bar{c}_{ii'},$$

   and the possible maximal agreement $\mathcal{M}_{\mathcal{V}\mathcal{V}'}$ between the two clusters:

$$\mathcal{M}_{\mathcal{V}\mathcal{V}'} = \sum_{i \in \mathcal{V}} \sum_{i' \in \mathcal{V}'} \mathcal{M}_{ii'}.$$

[4] This number is not fixed a priori, but will be discovered automatically during the first process

We will, then, merge the clusters, which have the best link (higher strict positive value). This must be carried out as long as there is a possibility to improve the criterion $C'(X)$.

2. **Transferring an observation from a cluster to another one**. When no cluster's merging is possible, we take the observations of each cluster and compute the link $\mathcal{L}_{i\mathcal{V}}$ (expression 6) of each observation O_i with the other clusters $\mathcal{V}$. If an observation has a better link with another cluster than its own, then this observation is transferred from its own cluster to this new cluster. This will be carried out, as long as improvement of the criterion $C'(X)$ occurs.

When, no observation's transfer is possible, we turn back to the merging step to see whether it is possible to improve the Condorcet's criterion by merging other clusters. These four steps will be applied, until no more improvements of the criterion occurred.

2.2.1 Illustrative example

Let us suppose that the studied population $\mathcal{P}$ is composed of seven observations $(O_1, O_2, \ldots, O_7)$ which have three qualitative variables $\left(V^1, V^2, V^3\right)$ were measured. The data set is presented in table 1.

	V^1	V^2	V^3
O_1	1	1	1
O_2	1	1	1
O_3	1	2	2
O_4	2	2	2
O_5	2	2	2
O_6	3	2	3
O_7	3	3	3

Table 1. Data set

After transformation of the three qualitative variables into their relational matrix representations, and after summing up those matrices, we obtain the Condorcet's global matrix C represented in table 2

	O_1	O_2	O_3	O_4	O_5	O_6	O_7
O_1	3	3	1	0	0	0	0
O_2	3	3	1	0	0	0	0
O_3	1	1	3	2	2	1	0
O_4	0	0	2	3	3	1	0
O_5	0	0	2	3	3	1	0
O_6	0	0	1	1	1	3	2
O_7	0	0	0	0	0	2	3

Table 2. Condorcet's global matrix C.

As the number of variables measured on this population is equal to three, it represents also the "maximum possible similarity" that can occur between two observations O_i and $O_{i'}$. We

can then deduce, that the global dissimilarity between those observations is $\bar{c}_{ii'} = 3 - c_{ii'}$. The binary squared matrix X, representing the obtained final partition of population $\mathcal{P}$ has the following general term:

$$x_{ii'} = \begin{cases} 1 & \text{if } c_{ii'} \geq \bar{c}_{ii'} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Due to the transitivity constraints, the solution is not so trivial[5] because of the so called "Condorcet's effect" cf. (Marcotorchino & Michaud, 1978; 1982), but the proposed heuristic is able to take into account some of those constraints limitations and avoid getting untransitive solutions. Applying the heuristic to the example (see Table 3), one gets the following optimal solution:

- Cluster 1: O_1, O_2
- Cluster 2: O_3, O_4, O_5
- Cluster 3: O_6, O_7

The relational representation X of this partition is then:

	O_1	O_2	O_3	O_4	O_5	O_6	O_7
O_1	1	1	0	0	0	0	0
O_2	1	1	0	0	0	0	0
O_3	0	0	1	1	1	0	0
O_4	0	0	1	1	1	0	0
O_5	0	0	1	1	1	0	0
O_6	0	0	0	0	0	1	1
O_7	0	0	0	0	0	1	1

Table 3. Binary matrix representation X of the final partition .

The corresponding Condorcet's criterion value is: $\mathcal{C}(X) = 131$.

3. Special case of clustering mixed data: Global Fusion

A specific SOM (Self-Organizing Map) model has been developed for mixed data using the similar cost function as the model presented in Kohonen (2001); Lebbah et al. (2005). The model dedicated to binary and continuous data is called MTM (Mixed Topological Map). As with a traditional self-organizing map, we assume that the lattice $\mathcal{C}$ (map) has a discrete topology defined by an indirect graph. Usually, this graph is a regular grid in one or two dimensions. For each pair of cells (c,r) on the map, the distance $\delta(c,r)$ is defined as the length of the shortest chain linking cells r and c. Let $\mathcal{P} = \{O_i, i = 1..n\}$ the learning data set where each observation $O_i = (O_i^1, O_i^2, ..., O_i^k, ..., O_i^m)$ is made of two parts: continuous part $O_i^{r[.]} = (O_i^{r[1]}, O_i^{r[2]}, ..., O_i^{r[d_r]})$ ($O_i^{r[.]} \in \mathcal{R}^{d_r}$) and binary part $O_i^{b[.]} = (O_i^{b[1]}, O_i^{b[2]}, ..., O_i^{b[k]}, ..., x_i^{b[d_b]})$ where the kth component $O_i^{b[k]}$ is binary variable ($O_i^{b[k]} \in \beta = \{0,1\}$) such as each observation O_i is thus, a realization of a random variable which belongs to $\mathcal{R}^{d_r} \times \beta^{d_b}$. With these notations a particular observation $O_i = (O_i^{r[.]}, O_i^{b[.]})$ is

[5] Just applying the rule (9) could yield to untransitive solution.

a mixed of subvectors (continuous and binary variables) of dimension $m = d_r + d_b$.

Since for binary vectors the Euclidean distance is no more than the Hamming distance $\mathcal{H}$, then the Euclidean distance can be rewritten by:

$$||O - \mathbf{w}_c||^2 = ||O^{r[.]} - \mathbf{w}_c^{r[.]}||^2 + \mathcal{H}(O^{b[.]}, \mathbf{w}_c^{b[.]})$$

where $\mathcal{H}(O^{b[.]}, \mathbf{w}_c^{b[.]})$ the complement of global similarity between a binary part of an observation O and referent $\mathbf{w}_c^{b[.]}$).

Using this expression, the cost function of the traditional SOM algorithm, which is dedicated to mixed data can be expressed as:

$$\begin{aligned} \mathcal{G}(\phi, \mathcal{W}) \quad = \quad & \sum_{O_i \in \mathcal{P}} \sum_{r \in \mathcal{C}} \mathcal{K}(\delta(r, \phi(O_i)))||O_i^{r[.]} - \mathbf{w}_r^{r[.]}||^2 \\ & + \sum_{O_i \in \mathcal{P}} \sum_{r \in \mathcal{C}} \mathcal{K}(\delta(r, \phi(O_i)))\mathcal{H}(O_i^{b[.]}, \mathbf{w}_r^{b[.]}) \end{aligned} \tag{10}$$

Where ϕ assigns each observation O_i to a single cell in $\mathcal{C}$. $\mathcal{K}$ is a particular kernel function which is positive and symmetric ($\lim_{|y| \to \infty} \mathcal{K}(y) = 0$).

The first term is the classical cost function used by the Kohonen Batch algorithm Kohonen (2001), and the second term is the cost function used in BinBatch model Lebbah et al. (2000). The cost function (10), is minimized using an iterative process with two steps.

1. Assignment step, which leads to the use of the following assignment function:

$$\forall O,\ \phi(O) = \arg\min_c \left(||O^{r[.]} - \mathbf{w}_c^{r[.]}||^2 + \mathcal{H}(O^{b[.]}, \mathbf{w}_c^{b[.]}) \right)$$

2. Optimization step: It is easy to see that this two minimizations of both terms allow to define:
   - The continuous part $\mathbf{w}_c^{r[.]}$ of the referent vector $\mathbf{w}_c$ as the mean vector as:

$$\mathbf{w}_c^{r[.]} = \frac{\sum_{O_i \in \mathcal{P}} \mathcal{K}(\delta(c, \phi(O_i)))O_i^{r[.]}}{\sum_{O_i \in \mathcal{P}} \mathcal{K}(\delta(c, \phi(O_i)))},$$

   - The binary part $\mathbf{w}_c^{b[.]}$ of the referent vector $\mathbf{w}_c$ as the median center of the binary part of the observations $O_i^{b[.]} \in \mathcal{P}$ weihted by $\mathcal{K}(\delta(c, \phi(O_i)))$. Each component $\mathbf{w}_c^{b[.]} = (w_c^{b[1]}, ..., w_c^{b[k]}, ..., w_c^{b[d_b]})$ is then computed as follows:

$$w_c^{b[k]} = \begin{cases} 0 & \text{if } \left[\sum_{O_i \in \mathcal{P}} \mathcal{K}(\delta(c, \phi(O_i)))(1 - O_i^{b[k]})\right] \geq \\ & \left[\sum_{O_i \in \mathcal{P}} \mathcal{K}(\delta(c, \phi(O_i)))O_i^{b[k]}\right] \\ 1 & \text{otherwise} \end{cases},$$

4. Experimental evaluation

In the following, the RA is used as the clustering consensus/fusion based algorithm for categorical and mixed data. First, the original data set is divided into two sub-data sets: pure categorical data set and pure continuous data set. Next, existing well established clustering algorithms designed for different data types are employed to provide corresponding clusters. We can run many algorithms or the same with different parameter using the same data. Finally the clustering results are combined as categorical data set to provide a consensus single clustering.

As quality evaluation criterion we use purity index. However, when class labels are available for each observation, we can use purity measure to indicate the match between cluster labels and class labels. The purity assess clustering quality from 0 (worst) to 1 (best).

5. Relational analysis for clustering categorical dataset

We used our RA clustering technique to cluster textual database "20 Newsgroups", which is a reference, for benchmarks for the data analysis scientific and technical community. This database is composed of 19997 documents, stemming from 20 different forums and described by 145980 descriptors (or variables). A major characteristic of this database is its heterogeneity both in terms of size of the documents and in terms of their themes and styles citelemoine.

At the end of the clustering process, we obtain 330 clusters. These clusters were sorted out in decreasing of magnitude (their size) order. As an example, we give here the list of the 7 first bigest clusters. Each cluster is described by the words or expressions (descriptors) participating the most into its constitution

Cluster	Descriptors	Cardinal
1	game, team, player, hockey, season, playoff, fan, baseball, league, coach	1325
2	file, directory, program, window, FTP, archive, DOS, disk, server	1144
3	Government, right, law, constitution, weapon, citizen, president, gun, policy	1095
4	Car, engine, mile, tire, mileage, brake, dealer, wheel, auto, clutch	755
5	Clipper, encryption, key, chip, escrow, crypto, wire tap, algorithm, privacy, government	673
6	Drive, SCSI, IDE, disk, controller, ram, floppy, CD-ROM, jumper, software	628
7	Card, video, driver, ISA, monitor, bus, VGA, VLB, SVGA, graphics	579

Table 4. The first seven clusters of the final partition.

Interpretation attempt
We can observe, in view of the descriptors characterizing the clusters that:

- the cluster 1 is compound of documents which generally deal, with "sport",
- the cluster 2 is compound of documents which are mainly related to "software" in general,
- the cluster 3 is built up with documents which are concerned with "politics"("policy"),
- the cluster 4 gathers documents which deal, in general, with "motorcar",
- the cluster 5 is made up of documents dedicated to "encoding and data protection",
- the cluster 6 is compound of documents which deal, generally, with "computer hardware" and more particularly with the choice between IDE or SCSI, and finally,
- the cluster 7 gathers documents which are also concerned with "computer hardware" and more particularly with video material.

5.1 Artificial data sets for fusion

We illustrate the cluster consensus applications on two artificial data sets downloaded from *http://strehl.com/* and used by (Strehl & Ghosh, 2002). The first data set (2D2K) was artificially generated and contains 500 observations each of two 2-dimensional (2D) Gaussian clusters. The second data set (8D5K) contains 1000 observations from multivariate Gaussian distributions (200 observations each) in 8D space.

For this experiment we take several clustering results provided by Strehl in his website *http://strehl.com/*. The authors provide two simulations of clustering ensemble: (FDC, Exp1) Feature-distributed Clustering (ODC, Exp2): Object-distributed Clustering. Table 5 indicates different results provided by Strehl and Ghosh adding the result obtained with our consensus clustering technique RA in both experimentations. Our purpose through this comparison, is not to assert that our method is the best, but to show that RA method can obtain quite the same good results as the two previews ones, without making any arbitrary assumptions about the number of clusters to be found. Indeed, as shown in the table bellow, we can see that RA method give similar results and quite comparable to the ones obtained by both proposed techniques (FDC, ODC). The main difference between these three methods is that, unlike the two other methods, RA doses not require a priori knowledge of the number of clusters.

8D5K		RA
FDC (Exp1)	0.9970	0.9930
ODC (Exp2)	0.9480	0.9330

2D2K		RA
FDC (Exp1)	0.9440	0.9440
ODC(Exp2)	0.9680	0.9700

Table 5. Comparison of consensus clustering. FDC: Feature-Distributed Clustering; ODC: Object-Distributed Clustering; RA: Relational Analysis; Exp: Experimentation

5.2 Real data sets and fusion

We will use three data sets coming from UCI repository (Asuncion & Newman, 2007). These data are mixed, in the sense that they contain both numerical and categorical data. These data are decribed bellow.

Heart disease data set: this data set, which is D. Detrano's heart disease data set, was generated by the Clevlande Clinic. It consists in 303 observations, described by 6 numerical and 8 categorical variables. The observations are also classified into two classes: healthy class (buff) and with heart-disease class (sick).

Credit data set : The data set has 690 instances, each being described by 6 continuous and 9 categorical variables. The observations were classified into two classes, approved class and rejected class.

Handwritten data: this data set consists of the handwritten numerals ("0"−"9") extracted from a collection of Dutch utility maps. There are 200 samples of each digit such that there is a total of 2000 samples. Each sample is a 15×16 binary pixel image. The data set is represented as a 2000×240 binary data matrix. Each categorical variable is a pixel with two possible values "On=1" and "Off=0".

In the first experiment we simulate such clustering result by running several clustering algorithms, each one having access to only a restricted categorical or continuous variables. Thus, each clusters has a partial view of the observations. The clusters are found using subspaces and adapted clustering technique. In the consensus clustering, cluster labels are clustered using RA technique. In order to compare our result, we cluster the data using a dedicated Self-organizing map for mixed categorical and continuous data. This technique is titled Mixed Topological Map (MTM), and provide a small cluster organized as map (see section 3). Often we use hierarchical clustering to reduce the number of the clusters (Vesanto & Alhoniemi, 2000). The combining method is indicated by MTM+HC and the number of clusters between bracket.

The figures 1 and 2 show the comparative results in term of number of clusters and the purity index. As can be seen, the both figures indicate that RA provides the high scores when compared for the same number of clusters. Note in this case for the both data set we have, a priori, two classes, and the RA (2) provides high purity for this case. We note also that RA don't require two steps of clustering, comparing to the MTM and other clustering ensemble algorithms found in the literature which needs an agglomerative clustering technique to reduce the number of clusters. The only parameter needed by RA is the similarity threshold.

In this second experiment we use Handwritten data set. The purpose is to use RA as consensus clustering of several runs of the same clustering algorithm. In this case we simulate 16 cluster results obtained with Self-organizing map dedicated to categorical data and hierarchical clustering, using different parameters (Lebbah et al., 2005; Vesanto & Alhoniemi, 2000). We use 5 cluster results with purity score lower than 0.4, and four results lower than 0.72, and the rest results are between 0.74 and 0.76. Thus the RA consensus clustering provide a stable purity with 0.76.

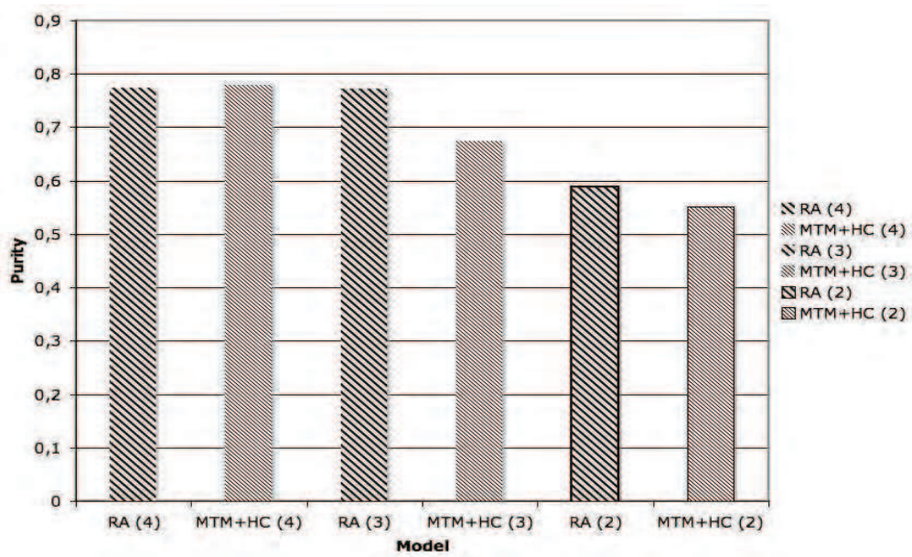

Fig. 1. Credit data set. Purity scores for consensus clustering. RA : Relational Analysis; MTM: Mixed Topological Map. HC: Hierarchical Clustering. The number between brackets indicates the number of clusters provided automatically

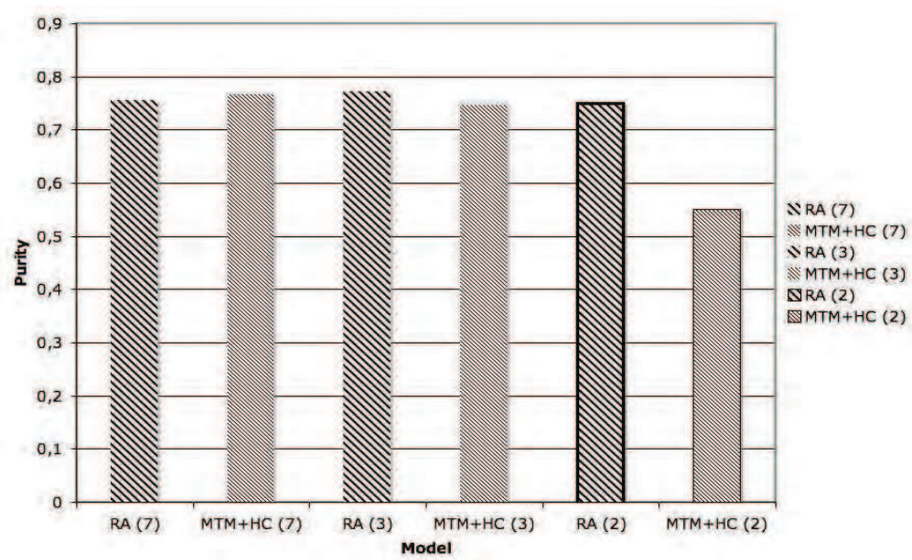

Fig. 2. Heart disease data set. Purity scores for consensus clustering. RA : Relational Analysis; MTM: Mixed Topological Map. HC: Hierarchical Clustering.The number between brackets indicates the number of clusters provided automatically

The figure 3 shows the distribution of each class of digit in all 15 consensus clusters. The figure 4 shows the best map obtained among the 16 maps used for consensus clustering. We visualize this figure in order to interpret the results of consensus. We note that RA grouped in a cluster numbered 12, 13, 15, the mix of digit 7, 9 and digit 5. It is clear to see on the map (Fig.4) that some figures such as "9" are written in the same way as "5" and "7". The same analysis could be done with the other clusters.

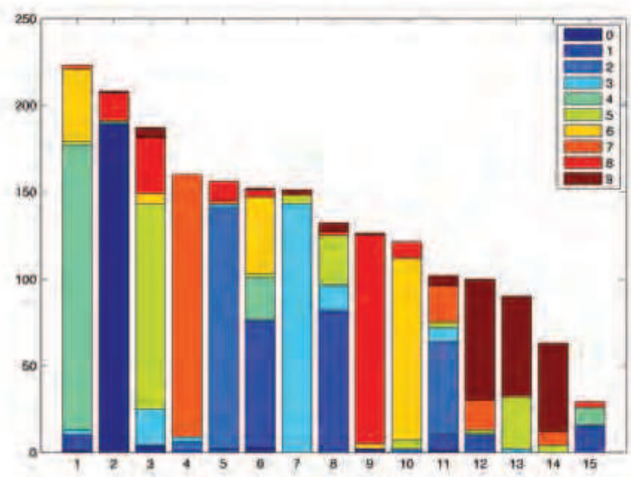

Fig. 3. Consensus clustering with RA. Each bar shows the distribution of each cluster.

6. Conclusions

In this chapter, we formally defined the problem of clustering and we presented an original and new approach of fusion/ensemble/consensus/aggregation clustering. The main idea was to find a clustering (or partition) of observations that represents the best consensus between several other clustering related to the same data set. The goal of the proposed algorithm is the improvement of confidence in cluster assignments by evaluating a history of cluster assignments for each observation. If we compare our algorithm (or method) to some recent clustering algorithm, we can assert that, unlike these new algorithms, our method is scalable, linear, in memory use and computational time and can handle data represented as observations cross attributes or as similarity matrix. Our clustering method handles missing values without replacing them by values that could be very far away from the true ones. It also contains a preprocessing module that, among other processings, can compute how discriminant are the attributes measured on the observations to be clustered. Finally we verified the intuitive appeal of the proposed approach and we studied the behavior of our algorithm on real and synthetic heterogeneous data sets. We observed that the proposed method increases performance as more as iterations of the process are performed. Another advantage of our method is that, neither do we need to re-process the data; nor do we need to fix the same cluster numbers for each application or clustering algorithm. In the future, we would like to perform a more detailed analysis involving huger data set and investigating the collaborative clustering.

Fig. 4. 16 × 16 map using MTM with only categorical data

7. References

Asuncion, A. & Newman, D. (2007). UCI machine learning repository, http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Azimi, J., Abdoos, M. & Analoui, M. (2007). A new efficient approach in clustering ensembles, *IDEAL, International Conference on Intelligent Data Engineering and Automated Learning*.

Benhadda, H. & Marcotorchino, F. (2007). L'analyse relationnelle pour la fouille de grandes bases de données., *Revue des Nouvelles Technologies de l'Information*, RNTI-A-2, Cépaduès, pp. 149–167.

Benhadda, H., Patino, J., , Corvee, E., Bremond, F. & Thonnat, M. (2007). Data mining on large video recordings, *Colloque V.S.S.T.2007 : Veille Strategique Scientifique & Technologique (21-25 Octobre) Marrakech* .

Condorcet, M. N. D. (1785). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix, *De l'imprimerie royale, Paris* .

Dudoit, S. & Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure, *Bioinformatics* **19**.

Frossyniotis, D. S., Pertselakis, M. & Stafylopatis, A. (2002). A multi-clustering fusion algorithm, *SETN '02: Proceedings of the Second Hellenic Conference on AI*, Springer-Verlag, London, UK, pp. 225–236.

Gionis, A., Mannila, H. & Tsaparas, P. (2007). Clustering aggregation, *ACM Trans. Knowl. Discov. Data* **1**(1): 4.

Kendall, M. G. & Smith, B. B. (1940). On the method of paired comparisons, *Biometrica* **31**: 324–345.

Kim, S. Y. & Lee, W. (2007). Ensemble clustering method based on the resampling similarity measure for gene expression, *Stat Methods Med Res* **16**: 539–564.

Kohonen, T. (2001). *Self-organizing Maps*, Springer Berlin.

Lebbah, M., Chazottes, A., Badran, F. & Thiria, S. (2005). Mixed topological map., *ESANN*, pp. 357–362.

Lebbah, M., Thiria, S. & Badran, F. (2000). Topological map for binary data, *Proceedings European Symposium on Artificial Neural Networks-ESANN 2000, Bruges, April 26-27-28*, pp. 267–272.

Marcotorchino, F. & Michaud, P. (1978). Optimisation en analyse ordinale des données, *Biometrica* **31**: 324–345.

Marcotorchino, F. & Michaud, P. (1982). Agrégation des similarités en classification automatique, *Revue de statistique appliquée* **30**(2): 21–44.

Minaei-Bidgoli, B., Topchy, A. & Punch, W. F. (2004). Ensembles of partitions via data resampling, *ITCC '04: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04) Volume 2*, IEEE Computer Society, Washington, DC, USA, p. 188.

Monti, S., Tamayo, P., Mesirov, J. & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data, *Mach. Learn.* **52**(1-2): 91–118.

Strehl, A. & Ghosh, J. (2002). Cluster ensembles – a knowledge reuse framework for combining multiple partitions, *Journal on Machine Learning Research (JMLR)* **3**: 583–617.

Topchy, A. P., Jain, A. K. & Punch, W. F. (2004). A mixture model for clustering ensembles, *SDM,proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA, April 22-24*.

Topchy, M.-A., Jain, F.-A. K. & Punch, W. (2005). Clustering ensembles: Models of consensus and weak partitions, *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(12): 1866–1881.

Vesanto, J. & Alhoniemi, E. (2000). Clustering of the self-organizing map, *Neural Networks, IEEE Transactions on* **11**(3): 586–600.

Machine Learning

Edited by Yagang Zhang

ISBN 978-953-307-033-9
Hard cover, 438 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

Machine learning techniques have the potential of alleviating the complexity of knowledge acquisition. This book presents today's state and development tendencies of machine learning. It is a multi-author book. Taking into account the large amount of knowledge about machine learning and practice presented in the book, it is divided into three major parts: Introduction, Machine Learning Theory and Applications. Part I focuses on the introduction to machine learning. The author also attempts to promote a new design of thinking machines and development philosophy. Considering the growing complexity and serious difficulties of information processing in machine learning, in Part II of the book, the theoretical foundations of machine learning are considered, and they mainly include self-organizing maps (SOMs), clustering, artificial neural networks, nonlinear control, fuzzy system and knowledge-based system (KBS). Part III contains selected applications of various machine learning approaches, from flight delays, network intrusion, immune system, ship design to CT and RNA target prediction. The book will be of interest to industrial engineers and scientists as well as academics who wish to pursue machine learning. The book is intended for both graduate and postgraduate students in fields such as computer science, cybernetics, system sciences, engineering, statistics, and social sciences, and as a reference for software professionals and practitioners.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Mustapha Lebbah, Younes Bennani, Nistor Grozavu and Hamid Benhadda (2010). Relational Analysis for Clustering Consensus, Machine Learning, Yagang Zhang (Ed.), ISBN: 978-953-307-033-9, InTech, Available from: http://www.intechopen.com/books/machine-learning/relational-analysis-for-clustering-consensus

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen